CANONICAL DECOMPOSITION OF CATENATION OF FACTORIAL LANGUAGES

A. E. FRID

Abstract. According to a previous result by S. V. Avgustinovich and the author, each factorial language admits a unique canonical decomposition to a catenation of factorial languages. In this paper, we analyze the appearance of the canonical decomposition of a catenation of two factorial languages whose canonical decompositions are given.

1. Introduction

This paper continues a research of decompositions of factorial languages started in [1, 2] and inspired by the field of language equations and algebraic operations on languages in general (see, e. g., [7, 8] and references therein). As the development of the theory shows, even language expressions where the only used operation is catenation prove very difficult to work with. It seems that nothing resembling the Makanin's algorithm for word equations (see, e. g., [4]) can appear for language equations with catenation. Even easiest questions tend to have very complicated answers. In particular, the maximal solution X of the commutation equation

$$LX = XL$$

may be arbitrarily complicated: as it was shown by Kunc [6], even if the language L is finite, the maximal language X commuting with it may be not recursively enumerable. This situation contrasts with that for words, since $xy = yx$ for some words x and y implies that $x = z^n$ and $y = z^m$ for some word z and $n, m \geq 0$.

In some sense, the problems of catenation of languages are due to the fact that a unique factorization theorem is not valid for it: as it was shown by Salomaa and Yu [9], even a finite unary language can admit several essentially different decompositions to a catenation of smaller languages, and an infinite language may have no decomposition to *prime* languages and all; here a language L is called *prime* if $L = L_1L_2$ implies that $L_1 = \{\lambda\}$, where λ is the empty word, and $L_2 = L$, or vice versa.

To avoid ambiguity of this kind, we restrict ourselves to *factorial* languages. This family is large and widely investigated since it includes, e. g., languages of factors of finite or infinite words and languages avoiding patterns (in the sense of [3]). We can also consider the factorial closure of an arbitrary language. Furthermore, the class of factorial languages is closed under taking catenation, unit, and intersection, and constitutes a monoid with respect to the catenation.

Decompositions of factorial languages to a catenation of factorial languages also may be several: for example, $a^*b^* = (a^* + b^*)b^* = a^*(a^* + b^*)$ (here and below

Frid, A. E., Catenation of factorial languages.

The work is supported by RFFI (grants 05-01-00364 and 06-01-00694).

(+) denotes unit). However, as it was proved in [1], we can define the notion of the *canonical* decomposition of a factorial language which always exists and is unique.

In this paper, we continue investigation of canonical decompositions of factorial languages and solve the following general problem: Given canonical decompositions of languages A and B, what is the canonical decomposition of their catenation AB?

Besides the self-dependent interest, the answer to this question may help to solve equations on factorial languages. Indeed, equal languages have equal canonical decompositions, and these canonical decompositions may be compared as words. So, techniques valid for words can be applied for them.

Thus, this paper may be considered as a description a tool helpful for solving equations on factorial languages.

2. Definitions and previous results

Let Σ be a finite alphabet, and $L \subseteq \Sigma^*$ be a language on it. A word $u \in \Sigma^*$ is called a *factor* of a word $v \in \Sigma^*$ if $v = sut$ for some (possibly empty) words s and t. The set of all factors of words of a language L is denoted by $\mathrm{Fac}(L)$. Clearly, $\mathrm{Fac}(\mathrm{Fac}(L)) = \mathrm{Fac}(L)$, so that $\mathrm{Fac}(L)$ may be called the *factorial closure* of L.

A language L is called *factorial* if $L = \mathrm{Fac}(L)$. In particular, each factorial language contains the empty word denoted by λ. In what follows, we consider only factorial languages.

The catenation of languages is an associative operation defined by

$$XY = \{xy | x \in X, y \in Y\}.$$

Clearly, languages constitute a monoid with respect to the catenation, and its unit is the language $\{\lambda\}$, where λ is the empty word. It is also clear that factorial languages form a submonoid of that monoid, since the catenation of two factorial languages is factorial.

A factorial language L is called *indecomposable* if $L = XY$ implies $L = X$ or $L = Y$ for all factorial languages X and Y.

Lemma 1. [1] *For each subalphabet $\Delta \subseteq \Sigma$, the language Δ^* is indecomposable.*

Other examples of indecomposable languages discussed in [1] include languages of factors of recurrent infinite words, etc.

A decomposition $L = L_1 \cdots L_n$ to factorial languages $L_1, \dots, L_n$ is called *minimal* if

- $L = \{\lambda\}$ implies $n = 1$ and $L_1 = \{\lambda\}$;
- If $L \neq \{\lambda\}$, then for $i = 1, \dots, n$ we have $L_i \neq \{\lambda\}$ and $L \neq L_1 \cdots L_{i-1} L_i' L_{i+1} \cdots L_n$ for any factorial language $L_i' \subsetneq L_i$.

A minimal decomposition to indecomposable factorial language is called *canonical.*

Theorem 1. [1] *A canonical decomposition of each factorial language L exists and is unique.*

In what follows, we shall denote the canonical decomposition of L by $\overline{L}$. Note that a canonical decomposition can be considered as a word on the alphabet $\mathcal{F}$ of all indecomposable factorial languages. In what follows, ($\doteq$) will denote equality of elements of $\mathcal{F}^*$; this notation will be used to compare canonical decompositions.

All examples of factorial languages we shall consider in this paper will be regular, just because regular languages are easy to deal with. Note that the factorial closure

of a regular language is always regular (which is a classical exercise). We have proved also

Theorem 2. [2] *If L is a regular factorial language, then all entries of $\overline{L}$ are also regular.*

3. Preliminary results

Suppose that we are given two factorial languages, A and B, on an alphabet Σ, and know their canonical decompositions $\overline{A}$ and $\overline{B}$. Our goal is to describe the canonical decomposition $\overline{AB}$, and the main result of the paper, Theorem 3, will give such a description. To state Theorem 3, we need to define two subalphabets of Σ, namely, Π and Δ.

For a factorial language L, let us define

$$\Pi(L) = \{a \in \Sigma | La \subseteq L\},$$

and

$$\Delta(L) = \{a \in \Sigma | aL \subseteq L\}.$$

Thus, if we take any word $u \in L$, we can extend it to the left by any word from $\Delta^*(L)$ and to the right by any word from $\Pi^*(L)$ to get a word from L. In other words, $L = \Delta^*(L)L\Pi^*(L)$, and $\Pi(L)$ and $\Delta(L)$ are defined as maximal languages with this property.

For the main result of this paper, we shall need to know the relationship between $\Pi(A)$ (further denoted by Π) and $\Delta(B)$ (further denoted by Δ). The following lemmas explain the meaning of these subalphabets. Note that analogues of Lemmas 2–5' were proved in [1], but the lemmas are reproved here both for the sake of completeness and of more precise wording.

Lemma 2. *If $\overline{L} \doteq L_1 \cdots L_k$, then $\Pi(L) = \Pi(L_k)$ and $\Delta(L) = \Delta(L_1)$.*

Proof. Let us prove the statement for $\Pi(L)$; the statement for $\Delta(L)$ is symmetric to it.

First, $\alpha \in \Pi(L_k)$ implies that $L_k\alpha \subseteq L_k$ and thus $L\alpha = L_1 \cdots L_k\alpha \subseteq L_1 \cdots L_k = L$; so, $\Pi(L_k) \subseteq \Pi(L)$.

On the other hand, $\alpha \in \Pi(L)$ means that $L_1 \cdots L_{k-1}v\alpha \subseteq L$ for all $v \in L_k$. Since L_k is a factor of the canonical decomposition of L, it cannot be contracted to a smaller factorial language L_k' such that $L_1 \cdots L_{k-1}L_k' = L$. It means that for each $v \in L_k \backslash \{\lambda\}$, there exists some word $wtv \in L$ such that $w \in L_1 \cdots L_{k-1}$, $tv \in L_k$, and w is the longest prefix of wtv belonging to $L_1 \cdots L_{k-1}$. Since tv is not the empty word, w is also the longest prefix from $L_1 \cdots L_{k-1}$ of the word $wtv\alpha \in L$. We see that $tv\alpha \in L_k$ and thus $v\alpha \in L_k$ since L_k is factorial. Moreover, by the same reason $\alpha \in L_k$, which means that $\lambda\alpha \in L_k$ and thus $L_k\alpha \subseteq L_k$. So, $\alpha \in \Pi(L)$ implies $\alpha \in \Pi(L_k)$, which was to be proved. □

Given a factorial language A and a subalphabet $\Delta \subseteq \Sigma$, let us define the factorial language $L_\Delta(A)$ =Fac$(A\backslash\Delta A)$. So, $L_\Delta(A)$ is the subset of A containing exactly words starting with letters from $\Sigma\backslash\Delta$ and their factors. Symmetrically, we define the subset $R_\Delta(A)$ of A containing exactly words which end with letters from $\Sigma\backslash\Delta$ and their factors: $R_\Delta(A)$ =Fac$(A\backslash A\Delta)$.

Lemma 3. *Let X and B be factorial languages on Σ. If there exists a factorial language A such that $X = AB$, then there exists a unique minimal one, and it is equal to $A' = R_{\Delta(B)}(A)$.*

Proof. First of all, let us prove that $A'B = X$. The $\subseteq$ inclusion is obvious: $A' \subseteq A$ and thus $A'B \subseteq AB = X$. To prove the $\supseteq$ inclusion, consider a word $x \in X$, and let b be its longest suffix from B: since $X = AB$, we have $x = ab$ for some word $a \in A$. Suppose that a ends with a symbol $\delta \in \Delta(B)$; then $\delta b \in B$ by the definition of $\Delta(B)$, and b is not the longest suffix of X belonging to B. A contradiction. Thus, $x = ab \in (A \backslash A\Delta(B))B \subseteq R_{\Delta(B)}(A)B = A'B$, and since x was an arbitrary element of X, the $\supseteq$ inclusion (and thus the equality $X = A'B$) is proved.

It remains to prove that $A' \subseteq Y$ for every factorial language Y such that $YB = X$. Let us consider an arbitrary non-empty word $a' \in A'$. Since $A' = R_{\Delta(B)}(A)$, the word a' is a factor of some word $sa't \in A \backslash A\Delta(B)$. Let the last letter of the word $sa't$ be equal to α; then $\alpha \in \Sigma\backslash\Delta$, and $a't = a''\alpha \in A$. So, $a'tB \subseteq AB = X = YB$.

For each $b \in B$, let us denote by $y(b)$ the longest prefix of $a'tb = a''\alpha b$ belonging to Y. Let the word b' be defined by the equality $a'tb = y(b)b'$; then $b' \in B$ since $a'tb \in YB$.

Clearly, if $y(b)$ is not shorter than a' for some $b \in B$, then its prefix a' belongs to Y (since Y is factorial), and this is what we need. But if $y(b)$ is shorter than a' for all $b \in B$, then each word b' contains αb as a suffix. So, $\alpha b \in B$ for all $b \in B$ (since B is factorial), and $\alpha \in \Delta(B)$ by the definition of $\Delta(B)$. A contradiction. So, $a' \in Y$ for all $a' \in A'$, and A' is indeed the minimal language such that $A'B = X$. □

Symmetrically, we can prove

Lemma 3' *Let X and A be factorial languages on Σ. If there exists a factorial language B such that $X = AB$, then there exists a unique minimal one, and it is equal to $B' = L_{\Pi(A)}(B)$.*

The following lemma is one of the main steps of the proof.

Lemma 4. *For each factorial languages A and B, we have*

$$\overline{AB} \doteq \overline{R_{\Delta(B)}(A)} \cdot \overline{L_{\Pi(R_{\Delta(B)}(A))}(B)} \doteq \overline{R_{\Delta(L_{\Pi(A)}(B))}(A)} \cdot \overline{L_{\Pi(A)}(B)}.$$

Proof. We shall prove the first equality; the second one can be proved symmetrically.

Let us denote $R_{\Delta(B)}(A) = A'$ and $L_{\Pi(R_{\Delta(B)}(A))}(B) = B''$. Due to Lemma 3, $A'B = AB$, and due to Lemma 3', $A'B'' = A'B$. So, $AB = A'B''$. Now note that all entries of the canonical decomposition of a language are indecomposable. So, to prove the required equality of canonical decompositions $\overline{AB} \doteq \overline{A'} \cdot \overline{B''}$, we must prove only that no entry of the canonical decompositions $\overline{A'}$ or $\overline{B''}$ can be decreased to get the same product.

Indeed, suppose we substituted an indecomposable entry of $\bar{(}A')$ by its proper factorial subset. Instead of A', we obtained its proper factorial subset A_1. Then $A_1B \subseteq AB$ since A' is the minimal factorial language such that $A'B = AB$. But $B'' \subseteq B$; so, $A_1B'' \subseteq A_1B \subsetneq AB$, and $A_1B'' \neq AB$.

Now suppose we substituted an indecomposable entry of $\overline{B''}$ by its proper factorial subset, and obtained a proper factorial subset B_1 of B''. Then $A'B_1 \neq A'B'' = AB$ since B'' is the minimal factorial set giving AB when catenated with A'.

So, no entry of $\overline{A'}$ or $\overline{B''}$ can be replaced by its proper subset without changing the result AB. The equality is proved. □

Lemma 5. *Let X and Y be factorial languages on Σ, and $\Delta \subset \Sigma$ be a subalphabet such that $Y \nsubseteq \Delta^*$. Then $R_\Delta(XY) = XR_\Delta(Y)$.*

PROOF. Consider a word $u \in XR_\Delta(Y)$. If $u \in X$, let us choose a symbol $y \in Y$ from $\Sigma \backslash \Delta$. Then $uy \in XY \backslash XY\Delta \subseteq R_\Delta(XY)$, and thus $u \in R_\Delta(XY)$. If $u \notin X$, then $u = xu'$, where x is the longest prefix of u belonging to X and $u' \in R_\Delta(Y)$ is a non-empty word. Let u'' be a word from $Y \backslash Y\Delta$ such that u' is its factor: $u'' = su't$ for some words s and t such that the last letter of t is from $\Sigma \backslash \Delta$. Then $u't \in Y \backslash Y\Delta$, and hence $ut = xu't \in XY \backslash XY\Delta \subseteq R_\Delta(XY)$. It follows that $u \in R_\Delta(XY)$, and the $\supseteq$ inclusion is proved.

To prove the $\subseteq$ inclusion, consider a word $u \in R_\Delta(XY)$. Let $u' = sut$ be a word from $XY \backslash XY\Delta$ whose factor is u, so that its last letter is from $\Sigma \backslash \Delta$. Then $ut \in XY \backslash XY\Delta$. Let $ut = xy$, where $x \in X$ and $y \in Y$; then $y \in Y \backslash Y\Delta$ and $ut \in X(Y \backslash Y\Delta)$. So, either $u \in X$, or $u = xy'$ for some prefix y' of y: since $y' \in R_\Delta(Y)$, in both cases we have $u \in XR_\Delta(Y)$, and the inclusion is proved. $\square$

Symmetrically, we prove

Lemma 5' *Let X and Y be factorial languages on Σ, and $\Pi \subset \Sigma$ be a subalphabet such that $X \nsubseteq \Pi^*$. Then $L_\Pi(XY) = L_\Pi(X)Y$.*

The following series of lemmas is also one of important parts of the main result.

Lemma 6. *Let X be a factorial language, $\Pi \subset \Sigma$ be a subalphabet, and $\Delta(X) \backslash \Pi \neq \emptyset$. Then $L_\Pi(X) = X$.*

PROOF. Let $\alpha \in \Sigma$ be a symbol from $\Delta(X) \backslash \Pi$; then each word u from X can be extended to $\alpha u \in X$ by the definition of $\Delta(X)$. So, $u \in$Fac$(\alpha u) \subset$Fac$(X \backslash \Pi X) = L_\Pi(X)$. Since u was chosen arbitrarily, and $L_\Pi(X) \subseteq X$, we get the equality: $L_\Pi(X) = X$. $\square$

The symmetric lemma is

Lemma 6' *Let X be a factorial language, $\Delta \subset \Sigma$ be a subalphabet, and $\Pi(X) \backslash \Delta \neq \emptyset$. Then $R_\Delta(X) = X$.*

Lemma 7. *For each factorial language X with $\overline{X} \doteq X_1 \cdots X_k$ we have*

$$\overline{L_{\Delta(X)}(X)} \doteq \begin{cases} X_2 \cdots X_k, & \text{if } X_1 = \Delta^*(X), \\ \overline{X}, & \text{otherwise.} \end{cases}$$

Symmetrically,

$$\overline{R_{\Pi(X)}(X)} \doteq \begin{cases} X_1 \cdots X_{k-1}, & \text{if } X_k = \Pi^*(X), \\ \overline{X}, & \text{otherwise.} \end{cases}$$

PROOF. We shall prove the first equality; the second one is symmetric. Let us denote $\Delta(X) = \Delta$.

Suppose first that $X_1 \neq \Delta^*$, that is, $X_1 \supsetneq \Delta^*$. Due to Lemma 5', $L_\Delta(X) = L_\Delta(X_1)X_2 \cdots X_k$. By the definitions, $X_1 = \Delta^* L_\Delta(X_1)$. But the language X_1 is indecomposable and is not equal to Δ^*, so, $X_1 = L_\Delta(X_1)$, and the equality $X = L_\Delta(X)$ (and thus $\overline{L_\Delta(X)} \doteq \overline{X}$) is proved.

Now suppose that $X_1 = \Delta^*$. Then $L_\Delta(X) = L_\Delta(X_2 \cdots X_k)$ by the definition of the operator L_Δ, since all elements of $X \backslash X_2 \cdots X_k$ cannot occur in $L_\Delta(X)$ anyway. Then, $L_\Delta(X_2) = X_2$ because otherwise we would have $X_1 X_2 = \Delta^* X_2 = \Delta^* Y$ for some $Y = L_\Delta(X_2) \subsetneq X_2$, contradicting to the minimality of the decomposition $\overline{X}$. So, due to Lemma 3', $L_\Delta(X) = L_\Delta(X_2 \cdots X_k) = L_\Delta(X_2)X_3 \cdots X_k = X_2 \cdots X_k$. The latter decomposition is minimal and thus canonical. $\square$

4. Main result

Theorem 3. *Let A and B be factorial languages with $\overline{A} \doteq A_1 \cdots A_k$ and $\overline{B} \doteq B_1 \cdots B_m$. Let us denote $\Pi = \Pi(A)$ and $\Delta = \Delta(B)$. Then the canonical decomposition of the catenation AB can be found as follows:*

(1) *If $\Delta \backslash \Pi \neq \emptyset$ and $\Pi \backslash \Delta \neq \emptyset$, then $\overline{AB} \doteq \overline{A} \cdot \overline{B}$.*
(2) *If $\Delta = \Pi$, and $A_k \neq \Delta^*$, $B_1 \neq \Delta^*$, then $\overline{AB} \doteq \overline{A} \cdot \overline{B}$.*
(3) *If $\Delta = \Pi$ and $A_k = \Delta^*$, then $\overline{AB} \doteq A_1 \cdots A_{k-1} \overline{B}$. Symmetrically, if $\Delta = \Pi$ and $B_1 = \Delta^*$, then $\overline{AB} \doteq \overline{A} B_2 \cdots B_m$.*
(4) *If $\Pi \subsetneq \Delta$, then $\overline{AB} \doteq \overline{R_\Delta(A)} \cdot \overline{B}$. Symmetrically, if $\Delta \subsetneq \Pi$, then $\overline{AB} \doteq \overline{A} \cdot \overline{L_\Pi(B)}$.*

Proof. Cases (1) and (4) are obtained directly by applying Lemmas 6 and 6' to the equality from Lemma 4. Case (2) is as well obtained by applying to Lemma 4 Lemma 7.

At last, in Case (3), if $A_k = \Delta^*$, we apply Lemmas 7 and 2 to get $\overline{L_\Delta(A)} \doteq A_1 \cdots A_{k-1}$ and $\Pi(L_\Delta(A)) = \Pi(A_{k-1})$. Assume that $\Pi(A_{k-1})$ includes Δ as a subset. Then $A_{k-1} = A_{k-1}\Delta^*$, and $A = A_1 \cdots A_{k-1}\Delta^* = A_1 \cdots A_{k-1}$, contradicting to the fact that $\overline{A} \doteq A_1 \cdots A_{k-1}\Delta^*$. So, $\Delta \backslash \Pi(A_{k-1}) \neq \emptyset$, and we apply Lemma 6 to get $L_{\Pi(A_{k-1})}(B) = B$. It remains to use Lemma 4 to get Case (3) of the Theorem. □

Corollary 1. *The canonical decomposition of AB either begins with $\overline{A}$, or ends with $\overline{B}$, so that only one of the languages A and B can give canonical factors of AB different from the canonical factors of the language itself.*

Example 1. If $A = \{a, b\}^*$ and $B = \{a, c\}^*$, then $\Pi(A) = \{a, b\}$, $\Delta(B) = \{a, c\}$, and the canonical decomposition of AB is just $\{a, b\}^* \cdot \{a, c\}^*$ (Case (1)).

Example 2. If $A =$Fac$\{a, ab\}^*$ and $B =$Fac$\{a, ac\}^*$, then $\Pi(A) = \Delta(B) = \{a\}$, and the canonical decomposition of AB is just Fac$\{a, ab\}^*$Fac$\{a, ac\}^*$ (Case (2)).

Here A is the language of all words on $\{a, b\}$ which do not contain two successive bs, and B is the language of all words on $\{a, c\}$ which do not contain two successive cs.

Example 3. If $A = a^*$ and $B =$Fac$\{a, ab\}^*$, then $\Pi = \Delta = \{a\}$, and $AB = B$ (Case (3)).

Example 4. Note that when $\Delta = \Pi$ and $A_k = B_1 = \Delta^*$, Case (3) may be applied in any of the two directions. For example, if $A = a^*b^*$ and $B = b^*a^*$, then $\overline{AB} \doteq a^* \cdot b^* \cdot a^*$, and it does not matter which of the occurrences of b^* was removed.

Before giving examples for Case (4), we will specify the form of the canonical decomposition of $A' = R_\Delta(A)$. Recall that A is a factorial language with the canonical decomposition $\overline{A} \doteq A_1 \cdots A_k$, and Δ is a subalphabet of Σ.

Let us define languages A'_i, $i = k, \dots, 1$, as obtained by the following iterative procedure: starting from $\Delta_k := \Delta$, we put for each i from k to 1

$$\begin{aligned} A'_i &= R_{\Delta_i}(A_i) \text{ and } \Delta_{i-1} = \Delta(A'_i), \text{ if } A_i \not\subseteq \Delta_i^*, \\ A'_i &= \{\lambda\} \text{ and } \Delta_{i-1} = \Delta_i, \text{ otherwise.} \end{aligned}$$

Lemma 8. *The canonical decomposition of $A' = R_\Delta(A)$ can be obtained by deleting extra $\{\lambda\}$ entries from the decomposition $A' = \overline{A'_1} \cdot \overline{A'_2} \cdots \overline{A'_k}$.*

PROOF. First of all, note that due to Lemma 5 applied iteratively, $A' = A_1 \cdots A_{k-1}A'_k = A_1 \cdots A_{k-2}A'_{k-1}A'_k = \ldots = A'_1 \cdots A'_k$. Some of the languages A'_i can be equal to $\{\lambda\}$; in particular, if $A \subseteq \Delta^*$, then $A' = \{\lambda\}$, as well as all its factors. However, if $A' \neq \{\lambda\}$, then we can canonically decompose factors A'_i not equal to $\{\lambda\}$ and erase the others.

Clearly, if we substitute any of canonical factors of A'_i by its proper subset, we get a new language $A''_i \subsetneq A'_i$. So, to prove the lemma, we should just show that $A' \neq A'_1 \cdots A'_{i-1}A''_i A'_{i+1} \cdots A'_k$ for any $A''_i \subsetneq A'_i$.

For all $i = 1, \ldots, k$, let us define $D_i = A'_1 \cdots A'_i$ and $E_{i-1} = A_1 \cdots A_{i-1}$. We also define $D_0 = \{\lambda\}$. Note that by the definition and Lemma 3, for all $i \geq 1$, D_i is the minimal language such that $D_i A'_{i+1} \cdots A'_k = A'$. So, it remains to prove only that $A'_i = A''_i$, where A''_i is the minimal language such that $D_{i-1}A''_i = D_i$. By Lemma 3', we have $A''_i = L_{\Pi(D_i)}(A'_i)$.

First, suppose that $D_{i-1} \neq E_{i-1}$. We knew that $D_i = D_{i-1}A'_i = E_{i-1}A'_i$, and D_{i-1} is the minimal language giving D_i when catenated with A'_i. So, by Corollary 1, in the canonical decomposition of D_i the factors corresponding to $\overline{A'_i}$ do not change, and $A'_i = A''_i$, which was to be proved.

Now suppose that $D_{i-1} = E_{i-1}$. Then $\Pi(D_{i-1}) = \Pi(E_{i-1}) = \Pi(A_{i-1})$. From now on, we denote this subalphabet just by Π'. We knew that A_i was equal to $L_{\Pi'}(A_i)$ since it was the minimal factorial language giving E_i when catenated with E_{i-1}. Assume by contrary that $A''_i = L_{\Pi'}(A'_i) \neq A'_i$.

Let us consider a word $u \in A'_i \backslash A''_i$. It does not belong to A''_i, which means that $su \in A'_i$ implies $su \in \Pi'\Sigma^*$ for all $s \in \Sigma^*$ (in particular, u starts with a letter from Π'). On the other hand, $u \in A'_i$, which means that $ut \in A_i \cap A_i(\Sigma\backslash\Delta_i)$ for some $t \in \Sigma^*$. By the definition, $ut \in A'_i$, and the set of non-empty left extensions of ut to elements of A_i is a subset of that for u:

$$\{s \in \Sigma^+ | sut \in A_i\} \subseteq \{s \in \Sigma^+ | su \in A'_i\} \subseteq \Pi'\Sigma^*.$$

Since we already know that $\lambda u = u \in \Pi'\Sigma^*$, we see that $ut \notin L_{\Pi'}(A_i)$. So, $A_i \neq L_{\Pi'}(A_i)$, contradicting to the fact that the decomposition $E_i = A_1 \cdots A_i$ was minimal. We have found a contradiction to the assumption that $A'_i \neq A''_i$.

So, $A'_i = A''_i$, and the decomposition obtained from $A' = A'_1 \cdots A'_k$ by deleting $\{\lambda\}$ entries is minimal, which was to be proved. □

To make the description complete, we state the symmetric lemma, for the case of $\Delta \subsetneq \Pi$. Let B be a factorial language with $\overline{B} \doteq B_1 \cdots B_m$ and Π be a subalphabet; we start from $\Pi_1 = \Pi$ and successively define for each $j = 1, \ldots, m$

$$\begin{aligned} B'_j &= L_{\Pi_j}(B_j) \text{ and } \Pi_{j+1} = \Pi(B'_j), \text{ if } B_j \not\subseteq \Pi_j^*, \\ B'_j &= \{\lambda\} \text{ and } \Pi_{j+1} = \Pi_j, \text{ otherwise.} \end{aligned}$$

The lemma symmetric to Lemma 8 is

Lemma 8' *The canonical decomposition of $\overline{B' = L_\Pi(B)}$ can be obtained by deleting $\{\lambda\}$ entries from the decomposition $B' = \overline{B'_1} \cdot \overline{B'_2} \cdots \overline{B'_m}$.*

The following easy example for Case (4) of Theorem 3 illustrates Lemma 8.

Example 5. The canonical decomposition of $A = (a^*b^*)^k + (b^*a^*)^k$ is $\overline{A} \doteq (a^* + b^*)^{2k}$ with $A_1 = \cdots = A_{2k} = (a^* + b^*)$ (here + denotes the unit). If we catenate it with $B = a^*$, we get $A'_{2k} = b^*$, $A'_{2k-1} = a^*$, and so on, and at last obtain $A' = (a^*b^*)^k$ and $\overline{AB} \doteq (a^* \cdot b^*)^k \cdot a^*$.

References

[1] S. V. Avgustinovich, A. E. Frid, *A unique decomposition theorem for factorial languages*, Internat. J. Algebra Comput. **15** (2005), 149–160.

[2] S. V. Avgustinovich, A. E. Frid, *Canonical decomposition of a regular factorial language*, in: D. Grigoriev, J. Harrison, E. Hirsch (Eds.), Computer Science - Theory and Applications, Springer, 2006 (LNCS 3967), 18–22.

[3] J. Cassaigne, *Unavoidable patterns*, in: in: M. Lothaire, Algebraic Combinatorics on Words, Cambridge Univ. Press, 2002. Pp. 111–134.

[4] V. Diekert, *Makanin's Algorithm*, in: M. Lothaire, Algebraic Combinatorics on Words, Cambridge Univ. Press, 2002. Pp. 387–442.

[5] J. Karhumäki, M. Latteux, I. Petre, *Commutation with codes*, Theoret. Comput. Sci. **340** (2005), 322-333.

[6] M. Kunc, *The power of commuting with finite sets of words*, in: Theoretical Aspects of Computer Science (STACS'05), Springer, 2005 (LNCS 3404), 569–580.

[7] M. Kunc, *Simple language equations*, Bull. EATCS 85 (2005), 81-102.

[8] A. Okhotin, *Decision problems for language equations with Boolean operations*, in: Automata, Languages and Programming, Spriger, 2003 (LNCS 2719), 239–251.

[9] A. Salomaa, S. Yu, *On the decomposition of finite languages*, in: Developments of Language Theory. Foundations, Applications, Perspectives, World Scientific, 2000, 22–31.

Anna E. Frid
Sobolev Institute of Mathematics,
pr. Koptyuga, 4,
630090, Novosibirsk, Russia
E-mail address: frid@math.nsc.ru